7-7-2020

Assessing Quality in Systematic Literature Reviews: A Study of Novice Rater Training

Sandra Acosta
Texas A&M University

Tiberio Garza
University of Nevada, Las Vegas, tiberio.garza@unlv.edu

Hsien-Yuan Hsu
University of Massachusetts Lowell

Repository Citation

Acosta, S., Garza, T., Hsu, H. (2020). Assessing Quality in Systematic Literature Reviews: A Study of Novice Rater Training. *SAGE Open, 10*(3), 1-11.
http://dx.doi.org/10.1177/2158244020939530

Original Research

SAGE Open
July-September 2020: 1–11

DOI: 10.1177/2158244020939530
journals.sagepub.com/home/sgo
SAGE

Assessing Quality in Systematic Literature Reviews: A Study of Novice Rater Training

Sandra Acosta[1], Tiberio Garza[2], Hsien-Yuan Hsu[3], and Patricia Goodson[1]

Abstract

This study investigated performance variability when graduate students critically appraised original studies from a systematic review. Fourteen doctoral students from different academic programs, with no systematic review experience, received training on the Methodological Quality Questionnaire (MQQ) rating scale. Participants were mostly male (71%) and non-native English speakers (79%). Each rater was randomly assigned one original study to independently assess using the MQQ. Their scores were compared to an expert rater. Statistical analysis comprised the following: percentage of agreement (POA), Kappa coefficient, and Kendall's tau-b correlation. On the completed MQQ rating scale, 43% of the novice raters had a POA of 78% or higher with the expert rater. From this case study, a guide for improving training on methodological quality assessment was developed. Benefits include the following: (a) developing and supporting critical reasoning as well as problem-solving skills and (b) increasing research skills and competencies in the systematic review process.

Keywords

Methodological Quality Questionnaire, training, systematic review, rater bias, higher education

Evidence-based practice and policy demand a systematic and rigorous approach for critically evaluating bodies of empirical studies on a specific topic. One platform for judging study quality is the systematic literature review, or systematic review, whose primary task is to evaluate "a field's knowledge claims while recognizing omissions, limits, and untested assumptions" (Rousseau et al., 2008, p. 479). In brief, systematic reviews synthesize evidence, identify gaps in the literature, and suggest productive lines of research that will increase knowledge and understanding, improve evidence-based decisions and choices, and in the long term support positive social change and better services (Andrews, 2005; Moja et al., 2005; Popay et al., 1998).

Evaluating a field's knowledge claims, in other words, conducting systematic reviews, requires judging the methodological quality of the studies producing those claims. Reviewers, therefore, are left with the tasks of establishing what constitutes "quality" in the context of each review, and which criteria, tools, and strategies to utilize for determining that quality. Having selected these quality measures, a key aspect, then, is how to enhance internal validity while minimizing error and bias when critically appraising original (primary) studies in a review.

Error and bias are sources of uncertainty in original research and in research syntheses as well. For systematic reviews, just as for original studies, error and bias can negatively impact data interpretation and seriously distort inferences drawn from study findings. In other words, error and bias are threats to validity and undermine the trustworthiness of the authors' conclusions, inferred either from an individual empirical study or multiple studies in a systematic review. Specifically, random error (variability) affects measurement precision; systematic error (bias) affects measurement accuracy (Dunn, 2004).

This study assessed whether structured training combined with practice using a rating scale can reduce error and bias when judging the quality of studies. To carry out this assessment, 14 doctoral students with little or no systematic review experience (i.e., novice reviewers) were trained to critically appraise the quality of individual original studies within a systematic review. The instrument employed for the appraisal process was a rating scale, the Methodological Quality Questionnaire (MQQ).

[1]Texas A&M University, College Station, USA
[2]University of Nevada, Las Vegas, USA
[3]University of Massachusetts Lowell, USA

Corresponding Author:
Sandra Acosta, Department of Educational Psychology, Texas A&M University, 107F Harrington Tower, MS 4225, College Station, TX 77843, USA.
Email: sacosta@email.tamu.edu

Given that assessing study quality is a core principle of systematic reviews (Petticrew, 2015), checklists and/or rating scales such as the MQQ are useful for the following: (a) diagnosing and assessing potential bias in the original study and (b) minimizing systematic reviewer bias during the coding and rating processes. The first item relates to the internal validity of the appraisal tool itself. The second item relates to raters' capacity for applying the conventions and decision rules as described in the MQQ criteria and rater training. In brief, instruments such as the MQQ allow us to assess study quality, make comparisons across studies in a body of research, and draw valid conclusions.

Systematic Review: Definition

The term systematic review at times has been misused and misappropriated (Bearman et al., 2012). Therefore, this study invokes the description of systematic review proffered by Moher et al. (2009; Preferred Reporting Items for Systematic Reviews and Meta-analyses [PRISMA]—the PRISMA definition was adopted from the Cochrane Collaboration):

> A systematic review is a review of a clearly formulated question that uses systematic and explicit methods to identify, select, and critically appraise relevant research and to collect and analyze data from the studies that are included in the review. Statistical methods (meta-analysis) may or may not be used to analyze and summarize the results of the included studies. (1)

In other words, "Systematic reviews are observational studies, in which prior studies are treated as sampling units and units of analysis" (Littell, 2008, p. 5).

Why investigate error and bias in systematic reviews? Primarily, to strengthen conclusions and reduce uncertainty (Littell et al., 2008; Petticrew & Roberts, 2006). Secondarily, although the number of published systematic reviews has increased in the social sciences, traditional narratives or nonsystematic literature reviews continue to hold sway (Bearman et al., 2012; Littell, 2013). Their value, together with their increased popularity, suggests that systematic reviews are invaluable tools for advancing evidence-based knowledge. Furthermore, systematic reviews have incorporated increasingly sophisticated methodologies. International research networks such as the Cochrane or Campbell Collaborations, for instance, continually develop/refine standards for conducting systematic reviews, regularly assemble expert evidence review teams, and produce/archive systematic reviews.

Systematic reviews matter in the social sciences, not merely because they afford a replicable method for advancing knowledge and clinical practice (Littell, 2013). Systematic reviews matter because, with these resources, review teams can produce research to promote and support social change in the areas of health, education, social welfare, criminal justice, and international development by investigating and disseminating high-quality evidence based on focal questions: "What helps? What harms? Based on what evidence (http://www.campbellcollaboration.org/)?" resulting in "Trusted evidence. Informed decisions. Better health (http://www.cochrane.org/)." Systematic reviews also matter to policymakers (positioning funding streams), administrators (allocating resources), and research teams (building hypotheses, developing explanatory theories, identifying "what works," and diagnosing gaps in the empirical evidence). At a more proximal level, systematic reviews matter to practitioners/clinicians seeking to increase their professional knowledge and effectiveness (see Cooper & Hedges, 2009; Gough et al., 2017; Littell, 2008; Oakley, 2003; Petticrew & Roberts, 2006; Saini & Scholonsky, 2012).

Systematic Reviews and Methodological Quality: Training Reviewers/Raters

Despite the increasing number of systematic reviews in doctoral dissertations, very little is known about training students, who are novice researchers, to assess study quality. For example, according to the ProQuest Dissertations and Theses Global database, between the years 1999 and 2016, 650 meta-analyses or systematic reviews were conducted for dissertations. Seventy percent were done between the years 2010 and 2016. This statistic highlights the growth of research synthesis among novice researchers.

Yet, a search of the research synthesis literature utilizing the electronic databases PubMed, Medline Complete, and EBSCO Academic Search Complete produced only two studies (McGuire et al., 1985; Oremus et al., 2012) on the topic of training reviewers, hereinafter referred to as *raters*, to assess individual study quality in systematic reviews. Both studies (a) recruited convenience samples from course cohorts, students with no previous experience assessing original study quality, and (b) investigated rater agreement (inter-rater reliability).

In the first study, McGuire et al. (1985) posed the question: Can methodological quality of original studies from systematic reviews/meta-analyses be appraised reliably? The authors (McGuire et al., 1985) proposed, then, to determine whether a weighting system employed in a meta-analysis approach could be applied to a random sample of 10 original studies from a published meta-analysis. The researchers asked two rater groups to rank order the studies by quality from best to worst using a blinded copy of the methods section from each original study. The methodology rater group comprised six advanced-level graduate students enrolled in a research methodology course. The substantive expert group comprised six nationally recognized researchers in the topic area of the meta-analysis. The student and expert groups exhibited little rater agreement: the Kendall's index for both groups was .29.

As potential remedies for increasing inter-rater reliability, McGuire et al. (1985) suggested future research explicitly define the methodological quality criteria and/or train the

raters. The authors (McGuire et al., 1985) concluded "agreement about methodological quality is not easily achieved" (5). They also urged others to continue investigating strategies for increasing rater agreement.

In the second study, Oremus et al. (2012) investigated inter-rater reliability and test–retest reliability of 10 inexperienced student raters (three undergraduate and seven graduate students). Similar to McGuire et al.'s (1985) study, student raters' initial agreement scores ranged from poor to fair. Nonetheless, there were notable research design differences between the two studies. First, unlike the McGuire et al. study (1985), in the study by Oremus et al. (2012), student raters received a 90-min training on two rating scales, the six-item Jadad Scale for randomized controlled trials (RCTs) and the Newcastle–Ottawa Scale (NOS) for observational studies, which students would subsequently use for rating their assigned original studies. Second, the Oremus et al. (2012) study employed a test–retest design. Student raters assessed the methodological quality of each assigned study twice. The second quality assessment occurred after a 2-month interval. Inter-rater reliability scores on the retest phase ranged from fair to excellent. Third, Oremus et al. (2012) did not compare the reliability between inexperienced student raters and experienced or expert raters. Concluding, the authors suggested a practice component ("pilot phase") following the rater training, where inexperienced raters might assess the methodological quality of a subsample of original studies (Oremus et al., 2012, p. 2).

The MQQ

The previous section argued for the importance of the systematic review as a knowledge development tool, along with the importance of investigating error and bias as sources of distorted interpretations. Here, the development and theoretical platform of the MQQ are presented as grounds for claims that the MQQ adequately measures original study quality. Accordingly, the MQQ is also a proper instrument for assessment of inter-rater reliability and rater bias.

The following instruments and standards informed the construction of the MQQ: the Jadad scale (Jadad et al., 1996), the methodological quality scoring system matrix (Goodson et al., 2006), and standards for reporting research published by both the American Psychological Association (APA, 2020) and the American Educational Research Association (AERA, 2006).

Three core assumptions underlie the MQQ's development. First, reporting standards for research (the MQQ criteria) used by major research associations (e.g., AERA) align with the dimensions of methodological quality. Second, the MQQ's scaling method (two-part dichotomous or binary rating scale [yes–no; agree–disagree]) quantitatively captures/summarizes empirical studies' methodological quality. Third, training raters to utilize the MQQ improves the measurement of methodological quality by minimizing rater differences, error, and bias; keeping rater response probabilities constant; and affording a platform for capacity building.

Validity theory is the theoretical polestar for the MQQ (see Kane, 2009, 2006; Messick, 1993). Messick defined validity as "integrative and evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions on test scores or other modes of assessment" (Messick, 1993, p. 13). The conceptual model of the MQQ depicts the relationship between the nine criteria as indicators of methodological quality and evidence validity. Hence, the degree of methodological quality, reported as criterion scores and total Methodological Quality Score (MQS), is a measure of support for the adequacy and appropriateness of the interpretation of the findings both at the micro level (individual original study) and the macro level (synthesis of the body of evidence—multiple studies).

The MQQ scoring system comprises nine criterion scores, with lower scores reflecting lower quality. The nine criterion scores are summed to produce one composite score—the MQS (see Supplemental Appendix A).

Each criterion is assessed on a dichotomous rating scale, consisting of two sequentially ordered tasks or items. Task 1 is a "yes–no" question (e.g., "Was the construct or phenomenon of interest theoretically or conceptually defined?") valued at 1 ("yes") or 0 ("no") points with no in-between or partial credit points. Task 2, completed only if the rater marked "yes" to the first task, is an "agree–disagree" extension statement (e.g., "The characteristics of the construct of interest or the relationship between the parameters were clearly defined.") valued at 2 ("agree") or 0 ("disagree") points. Tasks 1 and 2 include a short-statement rationale for supporting raters' responses. When raters mark "yes" or "agree" responses, they also note in the rationale section the page number from the original study supporting their response. The question and extension statement format provide transparency and minimize raters' inferences about each criterion when applied to a particular study. The scores on each of the tasks are summed to form the total points for each criterion. The total potential points for each criterion were "0" (lowest score), "1," or "3" (maximum score). The MQS, the composite score—the sum of all criteria scores—ranged from 0 to 27 points.

The nine MQQ criteria are as follows: (a) theoretical or conceptual definition of the focal variable(s) or construct(s); (b) operational definition of the focal variable(s) or construct(s); (c) research design; (d) sampling design; (e) sample; (f) reliability and validity evidence in quantitative studies or trustworthiness, credibility, and dependability in qualitative studies[1]; (g) data analysis; (h) implications for practice (topic-specific); and (i) implications for policy (topic-specific). Criteria eight and nine offer problem-specific flexibility. These criteria allow us to contextualize the MQQ by considering research standards related to the problem or topic and characteristics of the problem/topic being

investigated. In addition to the MQQ rating scale, we developed a structured training and checklist for guidance when completing the MQQ (See Supplemental Appendix A).

The rationale for selecting the nine criteria lies in their alignment with Journal Article Reporting Standards (JARS; APA, 2020) and Standards for Reporting on Empirical Social Science Research (APA, 2020). Each standard, vetted by expert committees from professional organizations, synthesizes theory-based criteria and best practices in reporting scientific inquiry. The MQQ criteria, drawn from these standards, provide a framework, via a questionnaire format, for reviewers to critically appraise the methodological quality of individual empirical studies in systematic reviews. Because study paradigm (e.g., qualitative, quantitative, and mixed methods) is not one of the nine criteria, all studies regardless of paradigm are given equal footing. An expert committee of researchers and systematic reviewers in higher education (disciplines: health education/promotion, social work, and educational psychology) vetted the MQQ's nine criteria and provided feedback throughout the development process.

To date, five systematic reviews (Acosta & Garza, 2011; Acosta et al., 2020; Huerta & Garza, 2019; Miller et al., 2018; Scott et al., 2018) have employed or adapted the MQQ for critically appraising study quality. Reliability evidence from three reviews consisted of percentage of rater agreement and a statistical measure of inter-rater agreement. In each systematic review, raters independently assessed and assigned MQS points to the studies. Percentage of rater initial agreement on the three reviews ranged from 81% to 85%. Although all three systematic review teams used consultations to determine final scores, after the systematic review published in 2011, the first author developed a training module for systematic review teams to employ when assessing study quality.

Purpose and Research Questions

To assess whether structured training can increase accuracy and minimize bias when judging original study quality, we designed a case study to (a) investigate performance variability when novice raters critically appraise study quality, employing a group of second year doctoral students who had no systematic review experience and (b) determine these raters' readiness for assessing study quality independently using the MQQ rating scale after participating in the training. To guide the present case study, two questions were posed:

1. After receiving training, will novice raters, using the MQQ rating scale, accurately identify different levels of methodological quality in studies drawn from a systematic review when compared to identification carried out by experts?
2. What do raters' agreements (i.e., hits) with the expert scores on individual criteria reveal about (a) the MQQ training effectiveness for novice reviewers and (b) areas for improving the MQQ training?

Method

Participants

This study used a convenience sample of doctoral students recruited through flyers and snowball sampling. The final sample consisted of 14 graduate students, from a university in the southwestern region of the United States. Participating doctoral students should have completed a minimum of 1 year in their doctoral programs but not be in the final semester before graduation. Participants represented the following programs/departments: architecture ($n = 1$), computer science ($n = 2$), education ($n = 3$), engineering ($n = 6$), health education ($n = 1$), and performance studies/liberal arts ($n = 1$). The sample was mostly male (71%; $n = 10$). English was the mother tongue of 21% ($n = 3$) of the sample; for 79% ($n = 11$), English was the second language. None of the participants had prior experience with systematic reviews.

Rating Instrument

Students applied the MQQ for critically appraising original study quality. Lower scores represented lower quality. See the previous section for a discussion of the development, composition/structure, and theoretical underpinnings of the MQQ.

Study Design

The training consisted of a two-step process, described in Table 1. In Step 1—training—1 week before the face-to-face training, the training facilitator e-mailed to participants the MQQ rating scale and a published study, which had been previously evaluated by an expert reviewer. The facilitator asked participants to use the MQQ instrument to assess the quality of the study before the training. Later, participants attended a 2-hr (face-to-face) meeting in which the facilitator and the participants compared their scores to the expert's scoring of the study. They discussed in detail their "hits" and "misses" (agreements and disagreements). In Step 2—data collection—participants were randomly assigned an original empirical study, independently assessed the study's quality using the MQQ, and completed a feedback survey.

Two-Step Training Process

Step 1. During the face-to-face meeting with study participants, the trainer (main author) explained the study purpose, distributed consent forms (to have their ratings analyzed and presented in this article), and gave participants the training

Table 1. MQQ Training Stages.

Stage	Activities		
Review: Foundational Knowledge	1. Agenda: Overview *of training and study* • Introductions—trainer and participants • Explanation of IRB consent forms and participant signatures • Description of the training plan and its components	2. Background: *Understanding systematic review process* • What is a systematic review? • How do systematic reviews, as a form of scientific inquiry, differ from other research reviews? • Why measure methodological quality? • What is the conceptual model for the MQQ rating scale? • How was the MQQ developed?	3. MQQ rating scale: *Criteria and scoring* • Operationalization of methodological quality—MQQ nine criteria and descriptors. • Describing the MQQ scoring system: criterion scores and total score (MQS) • Rating methodological quality using the MQQ: how to complete the MQQ rating scale
Practice Rating a Study: Formative Assessment	1. *Before Rating: Pre-appraisal* • Strategies for reading the practice research study efficiently to understand the issues, concepts, and context of the study topic (e.g., using word search, prioritizing the reading sequence of study report sections • Understanding the structure of a research study (report) vis-à-vis the nine criteria: where to find each MQQ criterion (e.g., C5 Sample) within the published research study report	2. *Application: Rating the study* • Review the MQQ checklist (guide for rating) • Procedures for rating the study individually on each criterion • Rate study on each criterion individually and write rationales and page numbers for talk-aloud and follow-up discussion	3. Diagnostics: Feedback • Talk-aloud strategy: using criterion rationales to make explicit participants' thinking to the trainer and other participants • Example of rationale— • Excerpt: "I don't see a specific research design named . . . first sentence says, 'research projects designed to document the impact . . . ' There must be some kind of design, but it's not named!" (C3 Research design) • Follow-up discussion: questions initiated by participants, probing questions from the trainer, participant reflections on rating process, and areas of confusion and clarity
Post-Training Evaluation: Summative Assessment	1. *Data Collection: Rating assigned study* • Raters randomly assigned to review a study with either high MQS, middle MQS, or low MQS • Instructions for rating assigned study • Instructions for submitting completed MQQ protocol and feedback survey	2. *Reactions: Feedback and reflections* • Participants' responses to feedback survey • Trainer's notes and/or reflection log	3. *Analysis: MQQ criterion and MQS scores* • Comparison of group MQS scores (high MQS, middle MQS, and low MQS) to expert score • Comparison of individual MQS and criteria scores to expert score

Note. MQQ = Methodological Quality Questionnaire; C = Criterion on the MQQ; IRB = Institutional Review Board; MQS = Methodological Quality Score.

materials. These comprised the following: a paper copy of the training module PowerPoint slides, the MQQ criteria in the form of a checklist, and the practice assessment activity (training stimuli)—a published empirical study previously appraised in a systematic review. The training, summarized in a PowerPoint file, consisted of three parts: Part 1, an overview of the MQQ: construction and development, 9-criteria rating scale, and scoring system; Part 2, practice: how to use the MQQ for rating the original study included in the training materials; and Part 3, instructions for rating the randomly assigned studies individually.

Step 2. Three original studies, drawn from a systematic review on the topic of high-stakes testing (Acosta et al., 2020), were categorized as high-, intermediate-, or low-level quality. These studies were previously rated by expert systematic reviewers and quantified in a total methodological quality score (MQS; see Acosta et al., 2020). Afterward, each participant was randomly assigned one study representing one of the three categories. The three selected studies included the following: (a) high-level quality: one study rated at 27 points (the maximum possible MQS); (b) intermediate-level quality: one study rated at 22 points, MQS at or above the median score (21 points; also, called the median threshold of the eligible original studies from the systematic review on high-stakes testing); and (c) low-level quality: one study rated at 16 points (MQS below the median threshold). Raters were identified by number (1–14) and study quality category (H = high,

I = intermediate, L = low; for example, R8I = rater 8 who assessed a study from the intermediate category).

The number of raters by quality category was as follows: high-level quality ($n = 4$), intermediate-level quality ($n = 5$), and low-level quality ($n = 5$). Raters were instructed to use the MQQ for independently assessing the methodological quality of their assigned study. Within 24 hr after completing the training, each participant received an e-mail with three documents: the MQQ, a feedback form about the study assessment process, and a copy of the published study to be rated. After encrypting the completed MQQ and feedback form, participants sent the completed MQQ coding sheet and the feedback form to the trainer via e-mail. Participants had 1 week from the day of face-to-face training to return the completed documents. No face-to-face meetings with the participants either in groups or individually occurred during Step 2.

Statistical Analysis

The criteria and methodological quality scores of the expert and novice reviewers were compared (see Table 2; here, the term "expert" or "expert score" refers to the final MQQ scores from the high-stakes systematic review). In that systematic review, two raters independently assessed each eligible study with discrepant scores resolved by a third rater. In essence, the expert score was the benchmark for comparing and evaluating the novice raters' MQQ scores. Note that the rating scores for each of nine criteria range from 0 to 3 and were in an ordinal scale. As a result, all experts and novices had nine scores corresponding to nine criteria. Simple statistics for measuring agreement and correlation were computed between novices' and expert's nine scores: percentage of agreement (POA), Kappa coefficient, and Kendall's tau-b correlation. The POA was computed by taking the number of agreements and dividing that by the total number of criteria (or the total number of possible agreements). The POA was provided to generally describe the agreement between raters (Kline, 2005).

Kappa coefficient was computed to inform the inter-rater reliability of categorical items between two raters (Kline, 2005). On the contrary, Kendall's tau-b correlation is a nonparametric statistic of the strength and direction of the relationship between two variables measured on at least an ordinal scale. Kendall's tau-b correlation was used rather than Spearman's correlation because of the small size of items in this study (Field, 2018). For the same reason, the statistical significance tests on Kendall's tau-b correlation were not used.

Note that both the Kappa coefficient and Kendall's tau-b correlation were computed by using a bootstrapping approach. The point estimates and corresponding confidence intervals are presented in Table 3. SPSS version 26 was utilized for data analysis.

Results

Critical Appraisal of Methodological Quality

Three statistics, including POA, Kappa coefficient, and Kendall's tau-b correlation, are presented in Table 3. Agreement patterns between expert and novice raters' scores clustered around study quality category (high-, intermediate-, and low-quality). Thus, as study quality decreases, POA also decreases. When assessing the range and median POA of all the raters in each category, this pattern is clear: high-quality category, POA scores ranged from 44% to 89%, with a POA median score of 78%. In the intermediate-quality category, POA scores ranged from 22% to 89%, with a median 67%. In the low-quality category, scores ranged 44% to 78%, with a median of 44%.

The Kappa coefficient could not be computed for the four raters in the high-quality category because the expert in this category assigned a score of 3 to all nine criteria, resulting in no variation. As a result, the Kappa coefficient (as well as Kendall's tau-b correlation coefficient) between this expert and novices were not computable. Therefore, only the POA for novices in the high-quality category were examined. Similarly, the novice R6I in the intermediate-quality category had a score of 3 for all nine criteria, and this novice's Kappa coefficient (and Kendall's tau-b correlation coefficient) was not computable either. Similar to what was found on the POA, novices in the intermediate-quality category had relatively higher Kappa coefficients than those in the low-quality category (range: –0.050 to 0.625, Mdn 0.197; range: –0.071 to 0.571, Mdn –0.023; respectively). One novice (R9I) in the intermediate-quality category and three novices (R10L, R12L, and R13L) in low-quality category had negative values of Kappa coefficient. (See Table 3 for Kappa coefficients.)

Regarding the Kendall's tau-b correlation coefficient, consistent with results of POA and Kappa coefficient, novices in the intermediate-quality category had relatively higher Kendall's tau-b correlation coefficients than those in the low-quality category (range: .058 to .592, Mdn 0.325; range: –.098 to .688, Mdn –.043; respectively). Three novices (R11L, R12L, and R13L) in low-quality category had negative values of Kendall's tau-b correlation. (See Table 3 for Kendall's tau-b correlation coefficients.)

No patterns of agreement emerged when assessing students' academic program or primary language (native or nonnative English speaker). Nonetheless, gender differences based on POA did surface. Females comprised 29% ($n = 4$) of the sample. Females' rates of agreement with expert scores (POA) ranged from 44% (education and engineering) to 89% (engineering), with a mean of 56% ($SD = 22\%$) and a median of 44%. Males' POA ranged from 22% (computer science) to 89% (architecture), with a mean of 63% ($SD = 21\%$) and a median of 72%.

Identical composite scores (MQS) among raters in the same quality group could produce different POAs. For

Table 2. A Comparison Between Expert Score and Novice Raters' Criteria and MQS.

Study assigned	Rater	Gender F/M	ESL	Academic program	Scores for Each criterion									MQS
					C1	C2	C3	C4	C5	C6	C7	C8	C9	
High quality	R1H	M	Yes	Architecture	1 2	1 2	1 2	1 2	1 2	0 0	1 2	1 2	1 2	24
	R2H	M	No	Education	1 2	1 2	1 2	1 2	1 2	1 0	1 2	1 2	1 0	23
	R3H	M	Yes	Engineering	1 2	1 2	1 2	1 0	1 0	0 0	1 2	1 0	0 0	15
	R4H	M	Yes	Engineering	1 2	1 2	1 2	1 2	1 2	0 0	1 2	1 2	0 0	21
	Expert Score				1 2	1 2	1 2	1 2	1 2	1 2	1 2	1 2	1 2	27
Intermediate quality	R5I	M	Yes	Engineering	1 2	1 2	1 2	1 0	1 2	1 0	1 2	1 0	0 0	18
	R6I	M	No	Liberal Arts	1 2	1 2	1 2	1 2	1 2	1 2	1 2	1 2	1 2	27
	R7I	F	Yes	Engineering	1 2	1 2	1 2	1 2	1 2	1 2	1 2	1 2	1 0	25
	R8I	M	Yes	Education	1 0	1 2	1 2	1 2	0 0	1 2	1 2	1 2	1 0	20
	R9I	M	No	Computer Science	1 2	1 2	1 0	1 0	0 0	0 0	0 0	1 0	0 0	9
	Expert Score				1 2	1 2	1 2	0 0	1 2	1 2	1 2	1 2	1 0	22
Low quality	R10L	F	Yes	Engineering	1 2	1 2	1 2	1 2	1 2	1 0	1 2	1 0	1 2	23
	R11L	F	Yes	Engineering	1 2	0 0	0 0	1 2	1 0	0 0	0 0	1 0	0 0	8
	R12L	F	Yes	Education	1 2	0 0	1 2	1 2	1 2	0 0	1 2	0 0	1 2	18
	R13L	M	Yes	Health	1 2	1 2	0 0	1 2	1 2	0 0	0 0	1 2	1 2	18
	R14L	M	Yes	Computer Sci.	1 2	1 2	1 2	0 0	0 0	0 0	1 2	1 2	1 2	18
	Expert Score				1 2	1 2	0 0	0 0	1 0	0 0	1 2	1 2	1 2	16

Note. MQS = Methodological Quality Scores; ESL = English as a Second Language; F = female; M = Male.

Table 3. Agreement and Correlation Between Expert and Novice Rating Scores of the MQQ.

Study assigned	Rater	Agreement with expert scores (%)	Kappa coefficient	Kendall's tau-b correlation
High quality	R1H[a]	88.89	—	—
	R2H[a]	77.78	—	—
	R3H[a]	44.44	—	—
	R4H[a]	77.78	—	—
Intermediate quality	R5I	55.56	0.143 (–0.105, 0.400)	.592 (.316, 1.000)
	R6I[a]	77.78	—	—
	R7I	88.89	0.625 (0.000, 1.000)	.548 (.147, 1.000)
	R8I	66.67	0.250 (–0.286, 0.786)	.058 (–.475, .800)
	R9I	22.22	−0.050 (–0.241, 0.100)	.101 (–.500, .592)
Low quality	R10L	44.44	−0.023 (–0.286, 0.280)	.111 (–.477, .756)
	R11L	44.44	0.167 (–0.400, 0.640)	−.043 (–.735, .604)
	R12L	44.44	−0.071 (–0.500, 0.500)	−.098 (–.632, .570)
	R13L	44.44	−0.071 (–0.500, 0.538)	−.098 (–.632, .617)
	R14L	77.78	0.571 (0.100, 1.000)	.688 (.175, 1.000)

Note. MQQ = Methodological Quality Questionnaire.

[a]Kappa coefficient and Kendall's tau-b correlation coefficient between novice and expert scores cannot be computed because either novice or expert had rating scores with no variation (e.g., expert score in the high-quality group and R6I score in intermediate-quality group; see Field, 2018; Kline, 2005; Miller & Lovler, 2016).

example, raters R12L, R13L, and R14L from the low-quality group had identical composite scores—18 MQS points (Table 2). Nonetheless, their POAs ranged from 44% to 77%. Upon closer inspection, although the total MQS was the same, the distribution of criteria scores across the MQQ was different. This distribution of criteria scores can be investigated using simple statistics, such as the Kendall's tau-b coefficient. Accordingly, the MQS provides important information about overall study quality but provides little other useful information. Conversely, partial agreement scores yield information about the patterns of agreement and disagreement, which are valuable for formative assessment.

In summary, the raters' MQQ scores provided two sources of information. First, 43% ($n = 6$) of the novice raters' POA with the expert score was 78% or higher, and 57% ($n = 8$) of novice raters' POA was 56% or higher. In other words, more

than 40% of the novice raters could identify and accurately assess 7 out of 9 (78%) methodological quality criteria when critically appraising original empirical studies.

Second, partial agreement on specific criteria provided insights into novice raters' understanding and/or confusion about underlying constructs. One example is Criteria Four—sampling design—where patterns of error occurred across the sample.

Novice Rater Training

Structured training effectiveness. Evidence that structured training was effective with the present study sample emerged in three ways. First, the number of "hits." After receiving training and then independently rating an original study (corresponding either to a high-, intermediate-, or low-quality level as assessed by an expert reviewer), novice reviewers' assessment of quality agreed with the expert's assessment. On the completed MQQ rating scale, 43% of novice raters' POA with the expert score was 78% or higher, and for 57% of novice raters, the POA was 56% or higher. In the high- and intermediate-quality groups, at least five raters' POA with the expert score was 78% or higher. In the low-quality study, there was only one rater with a POA of 78%, while other raters' scores were 44% ($n = 4$).

Second, even when reviewing outside one's field of training, raters in this sample provided evidence that they were able to adequately assess the quality of studies being reviewed. While 79% ($n = 11$) of raters represented academic programs other than education, all were asked to assess education studies. Errors on criteria 8 (implications for practice [in K-12 education]) and 9 (implications for policy [in K-12 education]) comprised only 23% of total errors across the nine MQQ criteria. In other words, novice raters from other academic programs could correctly assess methodological quality of original studies on topics outside their disciplines/fields.

Third, completing each MQQ rating scale accurately and completely according to the instructions is an important component of the assessment process. All raters completed the MQQ rating scale according to the instructions and scoring rules without error. Only one participant did not complete the rationales as instructed.

Areas for improving structured training. Improving structured training is an ongoing struggle to achieve a balance between essential and non-essential information, maximize learning, and support behavior change. Two areas for improving structured training and minimizing error and bias emerged from case study data and trainer's notes. First, the practice phase of the training should include a study exhibiting low quality and other studies representing different paradigms (e.g., qualitative and mixed methods).

Second, capacity building: the trainer should review and address potential misunderstandings and confusions about concepts such as the difference between sample and sampling design or between research design and sampling design. Only one rater in the low-quality category correctly rated Criterion 4 (sampling design). All other novice raters in the intermediate- and low-quality categories rated Criterion 4 incorrectly. Two excerpts from Criterion 4 (sampling design) rationales demonstrate novice raters' confusion about the difference between sample and sampling design; for instance, "I guess the sample is. . .children and graduate student teachers?" or "All sites used standardized tests." In both cases, the novice raters gave Criterion 4 the maximum score of 3 points, incorrectly affirming that the sampling design was described and described in enough detail to be replicable. In contrast, experts' score for Criterion 4 was 0 points.

Discussion

Findings from the present case study highlight the importance of systematically training novice raters to critically appraise original studies in systematic reviews. First, this investigation of error and bias was situated in the study appraisal process. It was argued that systematic reviews, as a methodological approach of research synthesis, matter because they are a source of quality evidence for policy makers and practitioners. The argument for systematic reviews was followed by explaining the development of the MQQ, a rating scale for assessing methodological quality in original studies. Finally, findings were provided from the present case study where a sample of 14 doctoral students from various disciplines completed training activities and afterward were randomly assigned one of three studies to assess independently. The three assessed studies represented one of the three methodological quality categories: high, intermediate, or low. Experts' assessment scores were used as the benchmark, against which novice raters' scores were compared.

Study authors expected the POA between expert and novice raters' scores would be high in the high- and low-quality groups. Nonetheless, agreement patterns mirrored study quality rankings: highest POA scores and POA median scores occurred in the high-quality category and lowest POA scores and POA median scores occurred in the low-quality category. No differences in POA resulted from academic discipline/field or ESL status. Given the brief MQQ training, findings from this study are encouraging and support training effectiveness, specifically building capacity by increasing raters' skill and knowledge about the critical appraisal process.

Developing cutoff points or scores to determine rater readiness for participating in systematic reviews is beyond the scope of this study. Nevertheless, insights can be extracted from the POA and partial agreement statistics about raters' capacity for accurately assessing study quality. For instance, a benchmark to indicate the need for further MQQ training or professional development/coursework in research methods

might be any POA below 70% and partial agreement scores that do not correlate significantly with the expert score.

Benefits of Systematic Review Training

What specific benefits or value does the systematic review process and explicit rater training hold for graduate students and faculty? Graduate students can advance their understanding of research topics in their discipline or field by conducting or participating in systematic literature reviews (Armitage & Keeble-Allen, 2008). In addition, they can contribute to the literature by employing systematic reviewing tools and skills (Jones, 2004; Owens et al., 2006; Perry & Hammond, 2002; Tuijin et al., 2012). Moreover, increasing students' understanding of the systematic reviewer–rater process promotes and reinforces a culture for inquiry and evidence-based practice (Minnie et al., 2010). Such a culture can foster novice reviewers' critical reasoning and problem-solving skills (Daigneault et al., 2014; Minnie et al., 2010; Sambunjak & Puljak, 2010) and support the development of academic socialization within a network of mentoring and research partnerships (given the team-based process of systematic reviews; Sambunjak & Puljak, 2010).

For faculty and educational researchers, the present case study and the training guide (see Supplemental Appendix B) serve as a heuristic for organizing and training systematic review teams, specifically the assessment of methodological quality. In addition, study authors anticipate that promoting the systematic review as a research method and facilitating the training process will help to raise awareness among current and future researchers about the importance of abiding by reporting standards and their value as a benchmark for quality reporting.

Contributions and Limitations

This article contributes to the research synthesis literature in two respects. First, the present case study addresses the need for rating scales, such as the MQQ, that demonstrate appropriate and acceptable psychometric properties. Second, this study acknowledges the void in the research synthesis literature on the topic of rater training and contributes to the literature through study findings. Moreover, unlike previous studies, the sample of graduate students were not recruited from a course cohort; they were recruited through flyers and snowball sampling technique to minimize the instructor effect. Furthermore, statistics on rater agreement included expert rater scores as opposed to calculating only the novice raters' scores. Third, the three-stage MQQ training activities are described in Table 1, and examples of simple statistical tests were provided for analyzing rater agreement. In addition, an MQQ training guide, based on lessons learned in this case study, provides procedures and protocols for conducting structured training that is explicit, transparent, and replicable. The goal of the training and its guide is to build rater capacity, as well as minimize rater error and bias during all stages of the critical appraisal process.

Despite these contributions, certain limitations must be recognized. One limitation relates to the MQQ itself. The MQQ uses published reports of original empirical studies for assessing methodological quality. For studies published in refereed journals, word limits may influence what is reported. To compensate for this limitation, the MQQ criteria were informed by APA and AERA reporting standards, which list and describe best practices for reporting research in journal articles.

Nonetheless, some aspects of quality might not be reported in publications and, therefore, not be captured in reviews. Ioannidis (2007) noted that study limitations were often not acknowledged in scientific literature. Thus, omissions or lack of clarity in reporting study limitations or weaknesses might result in the potential for readers to misinterpret threats to internal validity. Although the presence or absence of limitation statements is noted in systematic review findings, they are not quantified in the MQQ composite score.

Conclusion

Sound decision-making demands accumulating knowledge drawn from valid evidence, generated by applying robust knowledge development methods. Appraising original study quality via methodological quality protocols, checklists, and/or scales is one approach for appraising the linkage between research question, methods, findings, and interpretation. Hence, judging original studies' methodological quality is critical for establishing the validity of findings from research syntheses and, ultimately, their usefulness. Herein, study authors have argued for the importance of rater training and developed a training guide for use with novice raters. The intent is to encourage researchers and doctoral students to conduct and/or participate in systematic reviews by providing tools—the MQQ rating scale (Supplemental Appendix A), training guide (Supplemental Appendix B), and rater training (Table 1)—so that quality evidence might be generated, diffused, and applied in policy, practice, and decision-making.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Tiberio Garza https://orcid.org/0000-0002-5673-2011

Supplemental Material

Supplemental material for this article is available online.

Note

1. Saldaña (2011) explains the qualitative concepts of trustworthiness and credibility as components of the *auditing* process in qualitative research. They provide the transparency necessary for establishing the "*integrity* and *honesty*" of the study (Saldaña, 2011, p. 136). Trustworthiness refers to informing the reader about the research process. Credibility refers to the methodology employed to accurately reflect the story created from the data. Thus, trustworthiness subsumes credibility (Saldaña 2011).

References

Acosta, S., & Garza, T. (2011). The podcasting playbook: A typology of evidence-based podagogy for prek–12 classrooms with English language learners. *Research in the Schools*, *18*(2), 39–56.

Acosta, S., Garza, T., Hsu, H.-Y., Goodson, P., Padrón, Y., Goltz, H. H., & Johnston, A. (2020). The accountability culture: A systematic review of high-stakes testing and English learners in the United States during No Child Left Behind. *Educational Psychology Review*, *32*(2), 327–352. https://doi.org/10.1007/s10648-019-09511-2

American Educational Research Association. (2006). Standards for reporting on empirical social science research in AERA publications. *Educational Researcher*, *35*(6), 33–40. https://doi.org//10.3102/0013189X035006033

American Psychological Association. (2020). *Publication manual of the American Psychological Association* (7th ed.).

Andrews, R. (2005). The place of systematic reviews in education research. *British Journal of Educational Studies*, *53*(4), 399–416. https://doi.org//10.1111/j.1467-8527.2005.00303.x

Armitage, A., & Keeble-Allen, D. (2008). Undertaking a structured literature review or structuring a literature review: Tales from the field. *The Electronic Journal of Business Research Methods*, *6*(2), 103–114. www.ejbrm.com

Bearman, M., Smith, C. D., Carbone, A., Slade, S., Baik, C., Hughes-Warrington, M., & Neumann, D. L. (2012). Systematic review methodology in higher education. *Higher Education Research and Development*, *31*(5), 625–640. https://doi.org//10.1080/07294360.2012.702735

Cooper, H., & Hedges, L. V. (2009). Research synthesis as a scientific process. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 1–16). Russell Sage Foundation.

Daigneault, P.-M., Jacob, S., & Ouimet, M. (2014). Using systematic review methods within a Ph.D. dissertation in political science: Challenges and lessons learned from practice. *International Journal of Social Research Methodology*, *17*(3), 267–283. https://doi.org//10.1080/13645579.2012.730704

Dunn, G. (2004). *Statistical evaluation of measurement errors: Design and analysis of reliability studies* (2nd ed.). Hodder Arnold.

Field, A. (2018). *Discovering statistics using IBM SPSS statistics* (5th ed.). Sage.

Goodson, P., Buhi, E. R., & Dunsmore, S. C. (2006). Self-esteem and adolescent sexual behaviors, attitudes, and intentions: A systematic review. *Journal of Adolescent Health*, *38*(3), 310–319. https://doi.org//10.1016/j.jadohealth.2005.05.026

Gough, D., Oliver, S., & Thomas, J. (Eds.). (2017). *An introduction to systematic reviews* (2nd ed.). Sage.

Huerta, M., & Garza, T. (2019) Writing in science: Why, how, and for who? A systematic literature review of 20 years of intervention research (1996–2016). *Educational Psychology Review*, *31*(3), 533–570. https://doi.org//10.1007/s10648-019-09477-1

Ioannidis, J. P. A. (2007). Limitations are not properly acknowledged in the scientific literature. *Journal of Clinical Epidemiology*, *60*(4), 324–329. https://doi.org//10.1016/j.jclinepi.2006.09.011

Jadad, A. R., Moore, R. A., Carroll, D., Jenkinson, C., Reynolds, D. J. M., Gavaghan, D. J., & McQuay, H. J. (1996). Assessing the quality of reports of randomized clinical trials: Is blinding necessary? *Controlled Clinical Trials*, *17*(1), 1–12. https://doi.org//10.1016/0197-2456(95)00134-4

Jones, M. L. (2004). Application of systematic review methods to qualitative research: Practical issues. *Journal of Advanced Nursing*, *48*(3), 271–278.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). American Council on Education.

Kane, M. T. (2009). Validating the interpretations and uses of test scores. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 39–64). Information Age Publishing.

Kline, T. J. B. (2005). *Psychological testing: A practical approach to design and evaluation*. Sage.

Littell, J. (2008). Evidence-based or biased? The quality of published reviews of evidence-based practices. *Children and Youth Services Review*, *30*(11), 1299–1317. https://doi.org//10.1016/j.childyouth.2008.04.001

Littell, J. (2013). Guest editor's introduction to special issue: The science and practice of research synthesis. *Journal of the Society for Social Work and Research*, *4*(4), 292–299. https://doi.org//10.5243/jsswr.2013.19

Littell, J., Corcoran, J., & Pillai, V. (2008). *Systematic reviews and meta-analysis*. Oxford University Press.

McGuire, J., Bates, G. W., Dretzke, B. J., McGivern, J. E., Rembold, K. L., Seabold, D. R., Turpin, B. M., & Levin, J. R. (1985). Methodological quality as a component of meta-analysis. *Educational Psychologist*, *20*(1), 1–5. https://doi.org//10.1207/s15326985ep2001_1

Messick, S. (1993). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). The Oryx Press.

Miller, D. M., Scott, C. E., & McTigue, E. M. (2018). Writing in the secondary-level disciplines: A systematic review of context, cognition, and content. *Educational Psychology Review*, *30*(1), 83–120. https://doi.org//10.1007/s10648-016-9393-z

Miller, L. A., & Lovler, R. L. (2016). *Foundations of psychological testing: A practical approach* (5th ed.). Sage.

Minnie, K., vander Walt, C., Klopper, H., & Cummings, C. (2010, July). *Systematic or integrative review as research project for graduate study* [Paper presentation]. International Nursing Research Congress, Orlando, FL, United States. www.researchgate.net/profile/Karin_Minnie/publication/265189023_Systematic_or_integrative_review_as_research_project_for_graduate_study/

links/54b50cc20cf2318f0f971503/Systematic-or-integrative-review-as-research-project-for-graduate-study.pdf

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLOS Medicine*, *6*(7), e1000097. https://doi.org/10.1371/journal.pmed.1000097

Moja, L. P., Telaro, E., D'Amico, R., Moschetti, I., Coe, L., & Liberati, A. (2005). Assessment of methodological quality of primary studies by systematic reviews: Results of the metaquality study cross sectional study. *British Medical Journal*, *330*(7499), 1053–1055. http://www.jstor.org/stable/25459599

Oakley, A. (2003). Research evidence, knowledge management and educational practice: Early lessons from a systematic approach. *London Review of Education*, *1*(1), 21–33. https://doi.org//10.1080/1474846032000049107

Oremus, M., Oremus, C., Hall, G. B. C., McKinnon, M. C. E. C. T., & Cognition Systematic Review Team. (2012). Inter-rater and test-retest reliability of quality assessments by novice student raters using the Jadad and Newcastle-Ottawa scale. *BMJ Open*, *2*, e001368. https://doi.org//10.1136/bmjopen-2012-001368

Owens, E., Baez, M., & Tillman, S. (2006). Lessons learned: The student experience. *Contemporary Issues in Communication Science and Disorders*, *33*, 74–78.

Perry, A., & Hammond, N. (2002). Systematic reviews: The experiences of a PhD student. *Psychology Learning and Teaching*, *2*(1), 32–35.

Petticrew, M. (2015). Time to rethink the systematic review catechism? Moving from "what works" to "what happens." *Systematic Reviews*, *4*, Article 36. https://doi.org//10.1186/s13643-015-0027-1

Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences: A practical guide*. Blackwell Publishing.

Popay, J., Rogers, A., & Williams, G. (1998). Rationale and standards for the systematic review of qualitative literature in health services research. *Qualitative Health Research*, *8*(3), 341–351. https://doi.org//10.1177/104973239800800305

Rousseau, D. M., Manning, J., & Denyer, D. (2008). Evidence in management and organizational science: Assembling the field's full weight of scientific knowledge through syntheses. *The Academy of Management Annals*, *2*(1), 475–515. https://doi.org//.1080/19416520802211651

Saini, M., & Scholonsky, A. (2012). *Systematic synthesis of qualitative research*. Oxford University Press.

Saldaña, J. (2011). *Fundamentals of qualitative research*. Oxford University Press.

Sambunjak, D., & Puljak, L. (2010). Cochrane systematic review as a PhD thesis: An alternative with numerous advantages. *Biochemia Medica*, *20*(3), 319–126.

Scott, C. E., McTigue, E. M., Miller, D. M., & Washburn, E. K. (2018). The what, when, and how of preservice teachers and literacy across the disciplines: A systemtic literature review of nearly 50 years of research. *Teaching and Teacher Education*, *73*, 1–13. https://doi.org//10.1016/j.tate.2018.03.010 0742-051X

Tuijin, S., Janssens, F., Robben, P., & van den Bergh, H. (2012). Reducing interrater variability and improving health care: A meta-analytical review. *Journal of Evaluation in Clinical Practice*, *18*, 887–895. https://doi.org//10.1111/j.1365-2753.2011.01705.x